Exercice 3 : Algorithme « a priori », règles d'associations

Méthode : ● Lister tous les items simples (1-itemsets). On compte combien de fois chaque item apparaît, on calcule le support, c'est à dire la proportion de transactions contenant l'itemset :

$$\text{support}(X)=\frac{\text{nombre de transactions contenant } X}{\text{nombre total de transactions}}.$$

● On ne garde que ceux dont le support supérieur ou égale min_supp.

● Générer les candidats de taille 2 à partir des itemsets fréquents de taille 1. On calcule leur support, on garde ceux qui sont supérieur ou égal à min_supp (Fréquents).

● Continuer ainsi à partir des itemsets fréquents de taille k, on génère les candidats de taille $k+1$, et on élimine ceux dont tous les sous-ensembles ne sont pas fréquents.

● On s'arrête quand plus aucun nouvel itemset fréquent n'est trouvé.

Données : 6 Transactions : $T_1=\{A,B\}$; $T_2=\{A,B,C,D\}$; $T_3=\{A,B,D\}$; $T_4=\{A,B,D,F\}$; $T_5=\{A,C,D,E\}$; $T_6=\{B,C,D,F\}$.

Seuils : min_supp$=0{,}5$ et min_conf$=0{,}75$.

1. k-itemset fréquents à l'aide de l'Algorithme « a priori »

1-itemsets candidats à partir de la liste des transactions :

1-Itemset	Occurrence	Support	Décision
A	5	$5/6\approx0{,}8333$	Fréquent
B	5	$5/6\approx0{,}8333$	Fréquent
C	3	$3/6=0{,}5$	Fréquent
D	5	$5/6\approx0{,}8333$	Fréquent
E	1	$1/6\approx0{,}1667$	Non fréquent
F	2	$2/6\approx0{,}3333$	Non fréquent

On retient la liste des itemset dont le support est supérieur ou égal à 0,5 : $L_1=\{A,B,C,D\}$.

2-itemsets candidats à partir de L_1 :

2-Itemset	Occurrence	Support	Décision
AB	4	$4/6\approx0{,}6667$	Fréquent
AC	2	$2/6\approx0{,}3333$	Non fréquent
AD	4	$4/6\approx0{,}6667$	Fréquent
BC	2	$2/6\approx0{,}3333$	Non fréquent
BD	4	$4/6\approx0{,}6667$	Fréquent
CD	3	$2/6\approx0{,}3333$	Fréquent

On obtient : $L_2 = \{AB, AD, BD, CD\}$.

3-itemsets candidatsà partir de L_2 :

3-Itemset	Occurrences	Support	Décision
ABD	3	$3/6 = 0{,}5$	Fréquent
BCD	2	$2/6 \approx 0{,}3333$	Non fréquent

On obtient : $L_3 = \{ABD\}$.

Pas d'itemsets de taille 4.
Ensemble final des itemsets fréquents : $L_1 = \{A, B, C, D\}$; $L_2 = \{AB, AD, BD, CD\}$ et $L_3 = \{ABD\}$.

2. Règles d'association, confiance et lift

Rappels : • La **confiance** mesure à quel point une règle est fiable. Autrement dit, si X est présent, quelle est la probabilité que Y soit aussi présent. La formule est : $\text{confiance}(X \to Y) = \dfrac{\text{support}(X \cup Y)}{\text{support}(X)}$.

• Le **lift** mesure l'intérêt réel ou la force du lien entre X et Y. Il compare la fréquence de X et Y ensemble à ce qu'on aurait par hasard. La formule est : $\text{lift}(X \to Y) = \dfrac{\text{confiance}(X \to Y)}{\text{support}(Y)}$.

Méthode : Pour chaque itemset fréquent de taille supérieure ou égale à 2 :

- on divise L en toutes les combinaisons possibles : X→Y, où $X \cup Y = L$:
- on calcule la confiance de la règle :
- si $\text{confiance}(X \to Y) \geqslant \text{min\_conf}$ on garde la règle.

Puis, on peut calculer le lift de chaque règle pour juger de sa pertinence.

Rappel support :

Itemset	A	B	C	D	AB	AD	BD	CD	ABD
Support	$\frac{5}{6}$	$\frac{5}{6}$	$\frac{1}{2}$	$\frac{5}{6}$	$\frac{2}{3}$	$\frac{2}{3}$	$\frac{2}{3}$	$\frac{1}{2}$	$\frac{1}{2}$

Règles issues des 2-itemsets, $\text{min\_conf} = 0{,}75$:

Règle	Confiance	Décision	Lift
$A \to B$	$\frac{2}{3} \div \frac{5}{6} = \frac{4}{5} = 0{,}8$	Gardé	$\frac{4}{5} \div \frac{5}{6} = \frac{4}{5} = 0{,}96$
$B \to A$	0,8	Gardé	0,96
$A \to D$	0,8	Gardé	0,96
$D \to A$	0,8	Gardé	0,96
$B \to D$	0,8	Gardé	0,96
$D \to B$	0,8	Gardé	0,96
$C \to D$	1	Gardé	1,2
$D \to C$	0,6	Rejeté	

Règles issues des 3-itemsets, $\text{min\_conf} = 0{,}75$:

Règle	Confiance	Décision	Lift
$AB \rightarrow D$	$\frac{1}{2} \div \frac{2}{3} = \frac{3}{4} = 0{,}75$	Gardé	$\frac{3}{4} \div \frac{5}{6} = \frac{9}{10} = 0{,}9$
$D \rightarrow AB$	0,6	Rejeté	
$AD \rightarrow B$	0,75	Gardé	0,9
$B \rightarrow AD$	0,6	Rejeté	
$BD \rightarrow A$	0,75	Gardé	0,9
$A \rightarrow BD$	0,6	Rejeté	

Règles les plus pertinentes :

- **C → D** est la règle la plus pertinente car elle est à la fois toujours vraie (confiance = 1) et statistiquement liée ($\text{lift} > 1$). C'est la règle la plus significative (D est sur-représenté quand C est présent).
- Les autres règles ont une confiance élevée (0,75 à 0,8) mais des lifts inférieurs à 1 (0,9 à 0,96). (Si $\text{lift}(X \rightarrow Y) < 1$ cela signifie que X et Y apparaissent ensemble un peu moins souvent que prévu au hasard : donc la règle est peu informative, même si la confiance semble haute.

Interprétation : Imaginons un supermarché où C = céréales et D = lait.

Toutes les fois où un client achète des céréales (C), il achète aussi du lait (D) donc confiance est égale à 1. Le lift supérieur à 1 montre que cette occurrence n'est pas due au hasard, mais reflète un vrai lien d'achat.

Les autres produits (A, B, D) sont souvent achetés par tout le monde, donc ils apparaissent ensemble souvent, mais sans lien particulier.